Symmetric Tests and Confidence Intervals for Survival Probabilities and Quantiles of Censored Survival Data

Stuart Barber and Christopher Jennison

Department of Mathematical Sciences, University of Bath, Bath, BA2 7AY, U.K.

Summary
We describe existing tests and introduce two new tests concerning the value of a survival function. These tests may be used to construct a confidence interval for the survival probability at a given time or for a quantile of the survival distribution. Simulation studies show that error rates can differ substantially from their nominal values, particularly at survival probabilities close to 0 or 1. We recommend our new “constrained bootstrap” test for its good overall performance.

1. Introduction

We consider inference on the survivor function, \( S(t) = \Pr(T > t) \), from survival times subject to independent right censoring. We present size \( \alpha \) tests of a hypothesis \( S(t^*) = p^* \), noting that a family of such tests can be inverted to obtain \( 100(1 - \alpha)\% \) confidence intervals for the survival probability at a specific time or for a quantile of \( S(t) \): the confidence interval for \( S(t^*) \) is the set of values \( p \) for which \( S(t^*) = p \) is accepted and the confidence interval for \( S^{-1}(p^*) \) is the set of values \( t \) for which \( S(t) = p^* \) is accepted. The confidence interval for \( S(t^*) \) lies above \( p^* \) if \( H_0: S(t^*) = p^* \) is rejected in favour of \( S(t^*) > p^* \) and below \( p^* \) if \( H_0 \) is rejected in favour of \( S(t^*) < p^* \). Under \( H_0 \), these these two errors usually have differing practical consequences, thus it is appropriate to require a \( 100(1 - \alpha)\% \) confidence interval to lie entirely above or below the true survival probability with probability \( \alpha/2 \) each. This requires a two-tailed parent test with type I error \( \alpha \) divided equally between both tails.

Key words: Beta distribution; Bootstrap sampling; Censoring; Confidence interval; Greenwood’s formula; Hypothesis test; Kaplan-Meier estimate; Likelihood ratio test; Survival data.
Our objective is, therefore, to find a test of \( H_0: S(t^*) = p^* \), achieving type I error close to \( \alpha \), and rejecting \( H_0 \) in favour of \( S(t^*) > p^* \) and \( S(t^*) < p^* \) with probability approximately \( \alpha/2 \) each. Previous work in this area has concentrated on the median survival time; see Slud, Byar and Green (1984) and Strawderman, Parzen and Wells (1997). We have found additional difficulties in testing \( S(t^*) = p^* \) for values of \( p^* \) away from 0.5 and so seek a test which performs well over a broad range of \( p^* \) values.

We first consider tests based on the Kaplan-Meier estimate \( \hat{S}(t^*) \) which refer the distribution of \( \hat{S}(t^*) \) to normal, beta and bootstrap approximations. In all these cases, the estimated distribution of \( \hat{S}(t^*) \) under \( H_0: S(t^*) = p^* \) is greatly improved by making explicit use of the fact \( S(t^*) = p^* \). When testing whether a binomial probability is equal to \( p^* \), it is standard practice to take the null variance of \( \hat{p} \) from a sample of size \( n \) to be \( p^*(1 - p^*)/n \) rather than \( \hat{p}(1 - \hat{p})/n \), but analogous methods have often been overlooked in the related problem of testing a hypothesised survival probability for right-censored data. The two new tests which we introduce, based on a beta approximation and a “constrained” bootstrap, prove to be the most successful tests of this type.

Lastly, we consider Thomas and Grunkemeier’s (1975) likelihood ratio test. Despite being an early proposal, this test has received little attention and we are not aware of any small sample results beyond Thomas and Grunkemeier’s limited simulations. We have found this test to be superior to many more recent proposals and well worth further consideration.

2. Methods for testing \( H_0: S(t^*) = p^* \)

2.1 Normal approximations to the distribution of \( \hat{S}(t^*) \)

We assume failure times are from a distribution with survivor function \( S(t) \) and the observed data consist of times \( t_i \) and indicator variables \( \delta_i, i = 1, \ldots, n \), with \( \delta_i = 0 \) if \( t_i \) is an exact failure time and \( \delta_i = 1 \) if \( t_i \) is a censoring time. Let \( t_{(1)} < \ldots < t_{(r)} \) be the distinct ordered failure times and denote the number of subjects at risk just before \( t_{(j)} \)
by $n_j = \#\{i : t_i \geq t_{(j)}\}, j = 1, \ldots, r$. Defining $d_j$ to be the number of failures at time $t_{(j)}$, the Kaplan–Meier estimate of $S(t)$ (Kaplan and Meier, 1958) is

$$
\hat{S}(t) = \prod_{j : t_{(j)} \leq t} \frac{n_j - d_j}{n_j}, \quad 0 \leq t \leq \max_i \{t_i\}.
$$

Meier (1975) showed the distribution of $\hat{S}(t)$ to be approximately normal with mean $S(t)$ and variance consistently estimated by Greenwood’s (1926) formula

$$
\hat{\sigma}_G^2(t) = \{\hat{S}(t)\}^2 \sum_{j : t_{(j)} \leq t} \frac{d_j}{n_j(n_j - d_j)}.
$$

Hence, we obtain the “Greenwood test” which rejects $H_0$: $S(t^*) = p^*$ if $\hat{S}(t^*) \notin \{p^* \pm z_{\alpha/2} \hat{\sigma}_G(t^*)\}$, where $z_{\alpha/2}$ is the value exceeded with probability $\alpha/2$ by a $N(0,1)$ variate.

In approximating the distribution of $\hat{S}(t^*)$ under $H_0$: $S(t^*) = p^*$, it is reasonable to use the fact $S(t^*) = p^*$. The maximum likelihood estimate of $S(t)$ for $0 \leq t \leq t^*$ with probability masses only at observed failure times and subject to the constraint $\hat{S}(t^*) = p^*$ is

$$
\overline{S}(t; t^*, p^*) = \prod_{j : t_{(j)} \leq t} \frac{n_j + \lambda - d_j}{n_j + \lambda}, \quad 0 \leq t \leq t^*,
$$

(1)

where $\lambda$ is the unique solution to $\overline{S}(t; t^*, p^*) = p^*$. Using $\overline{S}(t; t^*, p^*)$ instead of $\hat{S}(t)$ in the derivation of Greenwood’s formula, we obtain the constrained estimate of $\text{var}\{\hat{S}(t^*)\}$,

$$
\hat{\sigma}_C^2(t^*; p^*) = (p^*)^2 \sum_{j : t_{(j)} \leq t} \frac{\hat{S}(t_{(j)}^-)}{n_j \overline{S}(t_{(j)}^-; t^*, p^*)} \overline{S}_{j}^{-1},
$$

where $\overline{S}_j = (n_j + \lambda - d_j)/(n_j + \lambda)$, $\overline{S}_j^- = 1 - \overline{S}_j$ and $\hat{S}(t^-)$ and $\overline{S}(t^-; t^*, p^*)$ denote values of $\hat{S}(t)$ and $\overline{S}(t; t^*, p^*)$ just before time $t$. This estimate is due to Thomas and Grunkemeier (1975). It is used by Jennison and Turnbull (1985), although the formula they state (p.620) is in error as $t_j$ occurs in two places where $t_j^-$ is needed. If $S(t^*) = p^*$, $\hat{\sigma}_C^2(t^*; p^*)$ and $\hat{\sigma}_G^2(t^*)$ are asymptotically equivalent, but $\hat{\sigma}_C^2$ has the advantage of reducing to the correct binomial variance, $p^*(1 - p^*)/n$, in the absence of censoring.
The “constrained variance” test rejects $H_0$: $S(t^*) = p^*$ if $\hat{S}(t^*) \notin \{p^* \pm z_{\alpha/2} \hat{\sigma}_C(t^*; p^*)\}$.

However, if no exact failures are observed before time $t^*$, $S(t^*; p^*)$ is undefined and we test $H_0$ directly but conservatively. Let $n_c$ denote the number of observations censored before $t^*$. Placing the failure probability $1 - p^*$ just before $t^*$, the number of subjects surviving past $t^*$, conditional on $n_c$ being censored before $t^*$, is $Bin(n - n_c, p^*)$. In the observed data all $n - n_c$ uncensored cases survived past $t^*$, and we reject $H_0$ if $(p^*)^{n-n_c} < \alpha/2$.

Figure 1 shows empirical type I error rates for the Greenwood and constrained variance tests in 5,000 simulated data sets. In each case, sample size was 100, failure times were exponentially distributed with mean 10, denoted $Exp(10)$, and censoring times were $Exp(50)$, so the probability of censoring was $1/6$. In this and subsequent figures, results are for $p^* = 0.1, 0.15, \ldots, 0.9$ and $t^*$ such that $S(t^*) = p^*$, the left and right bars of each pair represent rejection of $H_0$: $S(t^*) = p^*$ due to low and high values of $\hat{S}(t^*)$ respectively. The intended two-sided type I error is $\alpha = 0.05$ and the horizontal line at $\alpha/2 = 0.025$ indicates the target height for each bar. The empirical one-sided error rate from a sample of 5,000 will lie within the dotted lines at $0.025 \pm 0.006$ with probability 0.95 if the true rate is 0.025.

Figure 1 shows that the Greenwood test performs well when $S(t^*) = 0.5$ but poorly at high or low values of $S(t^*)$. The results are strongly asymmetric, most errors arising with $\hat{S}(t^*)$ too high for high $S(t^*)$ and too low for low $S(t^*)$. Error rates of the constrained variance test are closer to their target values over the whole range of $S(t^*)$, although there are still noticeable discrepancies for values of $S(t^*)$ above 0.7 or below 0.35. We have also simulated other survival and censoring distributions with similar results.

Other estimates of $\text{var}\{\hat{S}(t^*)\}$ have been proposed. We have used Peto et al.’s (1977) suggestion of $\hat{\sigma}_P^2(t) = \{\hat{S}(t)\}^2 \{1-\hat{S}(t)\}/h_k$ for $t(k) \leq t < t(k+1)$, but found this test to fare worse than the Greenwood test, especially at the high or low values of $S(t^*)$ for which this variance estimate is intended. Rothman (1978) approximates $\text{var}\{\hat{S}(t^*)\}$ under $H_0$: $S(t^*) = p^*$ by
\[ \hat{\sigma}_R^2(t^*; p^*) = \frac{p^*(1 - p^*)}{S(t^*) \{1 - \hat{S}(t^*)\}} \hat{\sigma}_G^2(t^*). \]

This estimate equals \( \hat{\sigma}_G^2(t^*) \) if \( \hat{S}(t^*) = p^* \) and, like the binomial variance for uncensored data, is proportional to \( p^*(1 - p^*) \). In fact, Rothman’s estimate gives numerical values close to the constrained variance estimate, \( \hat{\sigma}_G^2(t^*; p^*) \), and tests based on these two variance estimates have similar properties, the Rothman test doing slightly worse for \( S(t^*) \) below 0.2.

2.2 Other approximations to the distribution of \( \hat{S}(t^*) \)

Further study of simulation results helps explain the empirical error rates seen so far. Barber and Jennison (1998) demonstrate that the estimate \( \hat{\sigma}_G^2(t^*) \) varies systematically with \( \hat{S}(t^*) \), over-estimating \( \text{var}\{\hat{S}(t^*)\} \) when \( \hat{S}(t^*) \) is closer to 0.5 than \( S(t^*) \) and under-estimating when \( \hat{S}(t^*) \) is farther from 0.5, and this causes the asymmetric error rates in the two tails of the Greenwood test. In contrast \( \hat{\sigma}_C^2(t^*; p^*) \) and \( \hat{\sigma}_R^2(t^*; p^*) \) estimate \( \text{var}\{\hat{S}(t^*)\} \) accurately with no systematic dependence on \( \hat{S}(t^*) \). We conclude that any shortcomings of the constrained variance and Rothman tests are due to inadequacies of the normal approximation to the distribution of \( \hat{S}(t^*) \).

Anderson, Bernstein and Pike (1982) have suggested applying logit, arcsin or \( \log(-\log) \) transformations to \( \hat{S}(t^*) \) to produce a variable which is more nearly normally distributed. We found the third of these transformations, in which \( \log[-\log\{\hat{S}(t^*)\}] \) is treated as approximately normal with estimated variance \( \hat{\sigma}_G^2(t^*)/[\hat{S}(t^*) \log\{\hat{S}(t^*)\}]^2 \), to give the best results. However, Figure 2 shows this transformed test to be inferior to the constrained variance test. In addition, problems arise when \( \hat{S}(t^*) = 0 \) or 1 since \( \log[-\log\{\hat{S}(t^*)\}] \) and its estimated variance are undefined. Replacing \( \hat{\sigma}_G^2(t^*) \) by \( \hat{\sigma}_P^2(t^*) \) or \( \hat{\sigma}_C^2(t^*; p^*) \) in the variance estimate led to markedly worse results.

In our simulations we found the distribution of \( \hat{S}(t^*) \) to resemble a beta distribution, prompting us to use a beta approximation directly. A \( Beta(r, s) \), random variable has
expectation \( r/(r + s) \) and variance \( rs/\{(r + s)^2(r + s + 1)\} \). Equating these to the hypothesised probability \( p^* \) and constrained variance estimate \( \hat{\sigma}_C^2(t^*; p^*) \) gives the pair of equations \( r/(r + s) = p^* \) and \( rs/\{(r + s)^2(r + s + 1)\} = \hat{\sigma}_C^2(t^*; p^*) \). We use the solutions, \( r \) and \( s \), of these equations as the parameters of our beta approximation to the distribution of \( \hat{S}(t^*) \) and reject \( H_0: S(t^*) = p^* \) if \( \hat{S}(t^*) \not\in \{\beta(r, s, \alpha/2), \beta(r, s, 1 - \alpha/2)\} \), where \( \Pr\{\text{Beta}(r, s) < \beta(r, s, p) = p\} \). Abramowitz and Stegun (1972, p.945) provide the accurate approximation \( \beta\{r, s, p\} \approx r/(r + se^{2w}) \) where

\[
w = \frac{z_p \sqrt{h + \lambda}}{h} - \left( \frac{1}{2s - 1} - \frac{1}{2r - 1} \right) \left( \lambda + \frac{5}{6} - \frac{2}{3h} \right),
\]

\( h = 2\{(2r-1)^{-1} + (2s-1)^{-1}\}^{-1} \) and \( \lambda = (z_p^2 - 3)/6 \). The beta approximation is asymptotically equivalent to the normal approximation; see Bickel and Doksum (1977, p. 53). Its small sample accuracy is demonstrated in Figure 2. The cause of the one serious failing when \( S(t^*) = 0.1 \) lies in the lower tail of the distribution of \( S(t^*) \) which exhibits several spikes, corresponding to zero, one, and two uncensored survivors at time \( t^* \), and it is not surprising that the beta distribution should approximate this tail poorly.

2.3 Bootstrap confidence intervals and a new bootstrap test

The bootstrap offers an alternative approach to approximating the distribution of \( \hat{S}(t^*) \). Endpoints of Efron’s (1981) “percentile” confidence intervals are simply percentiles of the bootstrap sampling distribution of the quantity of interest. In testing \( H_0: S(t^*) = p^* \), bootstrap data sets are obtained by sampling with replacement from the observed pairs of survival times and censoring indicators, \( \{(t_i, \delta_i); i = 1, \ldots, n\} \). Kaplan-Meier estimates \( \hat{S}^B(t^*) \) of the survival function at time \( t^* \) for \( M \) bootstrap samples are written in order as \( \hat{S}^B_{[1]}(t^*) \leq \ldots \leq \hat{S}^B_{[M]}(t^*) \) and \( H_0 \) is rejected if \( p^* < \hat{S}^B_{[M\alpha/2]}(t^*) \) or \( p^* > \hat{S}^B_{[M(1-\alpha/2)]}(t^*) \).

If \( \hat{S}(t^*) - S(t^*) \) is approximately pivotal, its distribution can be approximated by that of the bootstrap quantity \( \hat{S}^B(t^*) - \hat{S}(t^*) \), leading to a different form of bootstrap confidence interval; see Hall (1992, Example 1.2) and Efron and Tibshirani (1993, Section
The related “pivotal” bootstrap test rejects $H_0$: $S(t^*) = p^*$ if $p^* < 2\hat{S}(t^*) - \hat{S}_{[M(1-\alpha/2)]}^{B}(t^*)$ or $p^* > 2\hat{S}(t^*) - \hat{S}_{[M\alpha/2]}^{B}(t^*)$.

The first two plots in Figure 3 show empirical type I error probabilities for percentile and pivotal bootstrap tests, using bootstrap samples of 1,000 replicates. The results are poor and qualitatively similar to those of the Greenwood test. Percentile-$t$ intervals (Hall, 1992, Section 3.2) performed poorly and bias corrected and accelerated percentile intervals (Efron, 1987) were inferior to the “constrained bootstrap” intervals which we now present.

The most successful parametric approximations to the distribution of $\hat{S}(t^*)$ under $H_0$: $S(t^*) = p^*$ make use of the information in $H_0$. We use this information in our new bootstrap method by sampling from a distribution satisfying $S(t^*) = p^*$ to provide a reference distribution for $\hat{S}(t^*)$ under $H_0$. We thus avoid the assumption that $\hat{S}(t^*) - S(t^*)$ is pivotal or can be transformed to a normality, or the need to correct inaccuracies in such assumptions.

In our constrained bootstrap we generate $n$ survival times, $f_i^B$, from $\mathcal{S}(t; t^*, p^*)$ as defined by (1) with potential censoring times, $c_i^B$, from $\hat{C}$, the Kaplan-Meier estimate of the censoring time distribution, failure being treated as right-censoring for this purpose. We calculate the Kaplan-Meier estimate, $\hat{S}^B(t; p^*)$, up to time $t^*$ for each bootstrap data-set, order the $M$ bootstrap estimates of $S(t^*)$ as $\hat{S}_{[1]}^B(t^*; p^*) \leq \ldots \leq \hat{S}_{[M]}^B(t^*; p^*)$, and reject $H_0$: $S(t^*) = p^*$ if $\hat{S}(t^*) < \hat{S}_{[M\alpha/2]}^B(t^*; p^*)$ or $\hat{S}(t^*) > \hat{S}_{[M(1-\alpha/2)]}^B(t^*; p^*)$. In fact, we right censor each $f_i^B$ just beyond $t^*$ since $\mathcal{S}(t; t^*, p^*)$ is only defined for $t \leq t^*$ but this suffices for determining $\hat{S}^B(t^*; p^*)$. If no failures are observed before $t^*$, $\mathcal{S}(t; t^*, p^*)$ is undefined and, as for the constrained variance test, we reject $H_0$ if $(p^*)^{(n-n_c)} < \alpha/2$, where $n_c$ is the number of observations censored before $t^*$. If all observations in a bootstrap data set are less than $t^*$ and the largest one is censored, $\hat{S}^B(t^*; p^*)$ is undefined and we take it to be zero.

Results for the constrained bootstrap test, shown in Figure 3, are clearly the best so far.
bootstrap samples, 1000 in our case, has little effect on the empirical rejection rate if the bootstrap provides a good approximation to the null distribution of the test statistic.

Despite the well-known connection between hypothesis tests and confidence intervals, little common ground is apparent between work on bootstrap confidence intervals and bootstrap hypothesis tests. Precedents in which simulations are conducted under a “closest element” of the null hypothesis are Silverman’s (1981, 1983) test for the number of modes of a probability density and the bootstrap sampling methods proposed by Zhang (1996, 1997). In Owen’s (1988) empirical likelihood method, the bootstrap sampling distribution satisfies a null hypothesis and Jennison’s (1992) bootstrap test for a hazard ratio, λ, uses bootstrap values of the score statistic generated under the hypothesis λ = λ0.

Coupling bootstrap samples for different tests can help in computing a constrained bootstrap confidence interval. Consider the general setting where bootstrap values \( T^B(\theta) \) of a test statistic are generated under the parameter value \( \theta \). Suppose the value \( t_{\text{obs}} \) has been observed and let \( g(\theta) = \text{pr}\{T^B(\theta) \geq t_{\text{obs}}\} \). If the distribution of \( T \), and so of \( T^B(\theta) \), is stochastically increasing in \( \theta \), a \( 100(1 - \alpha)\% \) bootstrap confidence interval for \( \theta \) is \((\theta_l, \theta_u)\) where \( g(\theta_l) = \alpha/2 \) and \( g(\theta_u) = 1 - \alpha/2 \). Finding \( \theta_l \) and \( \theta_u \) is complicated by the fact that simulation based estimates of \( g(\theta) \) may not be monotonic but this problem disappears if bootstrap realisations can be coupled so that the estimate of \( g(\theta) \) is monotonic in \( \theta \). When computing a constrained bootstrap confidence interval for \( S(t^*) \), such coupling is achieved by using a common sequence of uniform random variates for each hypothesised \( p^* \) in determining survival past the possible failure times \( t_{(1)}, \ldots, t_{(r)} \).

### 2.4 Thomas and Grunkemeier’s Likelihood Ratio test

Thomas and Grunkemeier (1975) derived the constrained estimate of \( S(t), \overline{S}(t; t^*, p^*) \), for use in constructing a likelihood ratio confidence interval for \( S(t^*) \). The equivalent test of \( H_0: S(t^*) = p^* \) rejects \( H_0 \) at the \( 100\alpha\% \) level if \(-2R > \chi^2_{1, \alpha}\) where \( R \) is the log-likelihood ratio.
of the observed data under (a) the constrained estimate and (b) the Kaplan-Meier estimate of \( S(t) \). Thomas and Grunkemeier show that

\[
R = \sum_{i : t_{(i)} \leq t^*} \left\{ (n_i - d_i) \log \left( 1 + \frac{\lambda}{n_i - d_i} \right) - n_i \log \left( 1 + \frac{\lambda}{n_i} \right) \right\}
\]

and outline a proof that \(-2R\) has asymptotically a \( \chi^2_1 \) distribution. Li (1995) gives a rigorous proof of this result, and Murphy (1995) further justifies the use of this and another extension of the likelihood ratio test to obtain confidence intervals for the cumulative hazard function.

The results in Figure 3 show Thomas and Grunkemeier’s test to perform well over the range of \( S(t^*) \). It attains the nominal error probabilities in each tail more accurately than the beta test and only a little less accurately than the constrained bootstrap test. Some asymmetry is apparent, rejection of \( H_0 \) tending to occur in favour of higher survival when \( S(t^*) \) is high and in favour of lower survival when \( S(t^*) \) is low.

3. Further simulations and conclusions

We have applied the constrained variance, beta, constrained bootstrap and likelihood ratio tests to more demanding examples with higher censoring and smaller sample sizes. Figure 4 shows empirical error rates from 5,000 simulations using \( Exp(10) \) survival times and \( Exp(25) \) censoring times. The sample size is 100 for results on the left and 50 for those on the right. As sample size is decreased or censoring increased, it becomes increasingly likely that \( \max_i \{ t_i \} < t^* \) and the Kaplan-Meier estimate of \( S(t^*) \) is undefined. The results presented are conditional error rates given that \( \hat{S}(t^*) \) is defined but we have omitted completely cases for which \( \hat{S}(t^*) \) was undefined in 10 or more of the 5,000 data sets.

All four tests continue to do well at values of \( S(t^*) \) between about 0.7 and 0.4. The performances of the constrained variance and beta test suffer most at high and low values of \( S(t^*) \), while there is not a great deal to choose between the constrained bootstrap and Thomas & Grunkemeier’s likelihood ratio test which both fare better. We have seen similar results in further simulations not reported in detail here.
Discreteness is apparent when $S(t^*)$ is high and there is significant probability of zero, one or two failures by time $t^*$. The constrained bootstrap only rejects $H_0$ if $\hat{S}(t^*)$ is strictly greater than or strictly less than a proportion $\alpha/2$ of the bootstrap replicates. Consequently, this test errs on the side of conservatism where the continuous approximations used by the other three tests can lead to either liberal or conservative error rates.

We recommend the constrained bootstrap test for its accuracy and its conservative treatment of the discreteness problem. However, this method requires the greatest amount of computation and, if sufficiently many uncensored failure times are available, the constrained variance, beta or likelihood ratio tests may be preferred for their computational simplicity.

ACKNOWLEDGEMENT
S. Barber was supported by an EPSRC studentship. The authors thank Dr Andrew Wood for discussions concerning bootstrap methods and referees for their useful comments.

REFERENCES


Figure 1: Empirical error rates from 5,000 simulations of Greenwood and constrained variance tests with $n = 100$, survival times $\sim \text{Exp}(10)$ and censoring times $\sim \text{Exp}(50)$.

Figure 2: Empirical error rates from 5,000 simulations of tests based on the log(-log) transform and the beta approximation of $\hat{S}(t^*)$. Here $n = 100$, survival times $\sim \text{Exp}(10)$ and censoring times $\sim \text{Exp}(50)$. 

13
Figure 3: Empirical error rates from 5,000 simulations of the “percentile”, “pivotal” and “constrained” bootstrap tests and Thomas and Grunkemeier’s likelihood ratio test. Here $n = 100$, survival times $\sim \text{Exp}(10)$ and censoring times $\sim \text{Exp}(50)$
Figure 4: Error rates from 5,000 simulations of four tests with survival times $\sim \text{Exp}(10)$ and censoring times $\sim \text{Exp}(25)$. Results on the left are for $n = 100$ and those on the right for $n = 50$. Cases where $\hat{S}(t^*)$ was undefined in more than 10 trials out of 5,000 have been omitted.